

## Introduction

### Problem

In biomedical domain, researchers often want to query over multiple data sets, such as PubMed, UniProt, internal lab data, etc. These data sources are often in heterogeneous format and so make query processing and analysis very difficult.

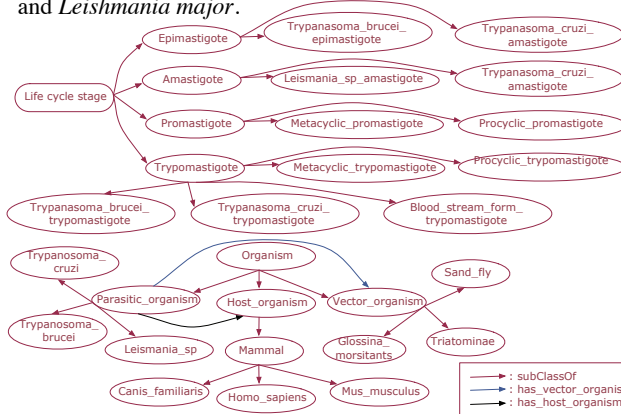
### Our solution

★ We propose an ontology-driven approach to develop a semantically integrated parasite knowledge repository (PKR) for querying over heterogeneous data. PKR along with an intuitive graphical query processing tool, Cuebee<sup>1</sup> will enable biological researchers to construct complex queries without any programming skills.

★ Both Parasite Experiment Ontology<sup>1</sup> (PEO) and Parasite Lifecycle Ontology<sup>1</sup> (PLO) have been developed and released through National Center for Biomedical Ontology (NCBO)<sup>2</sup>.

## Parasite Lifecycle Ontology

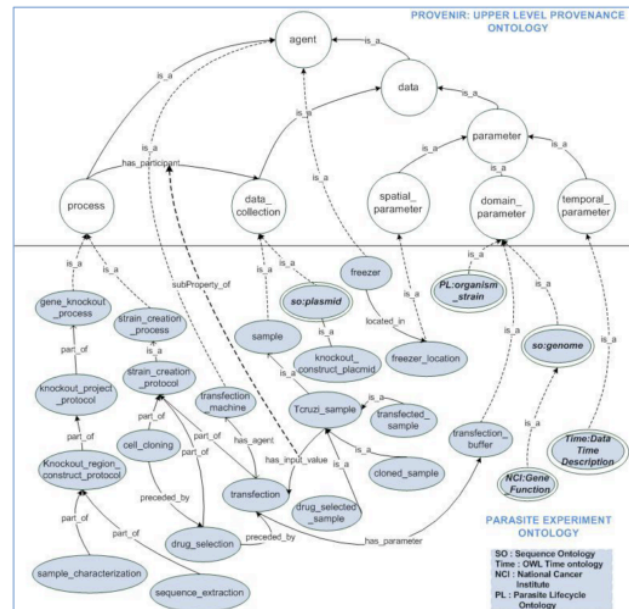
PLO models the lifecycle stage details of *Trypanosoma cruzi* and two related kinetoplastids, *Trypanosoma brucei* and *Leishmania major*.



## Parasite Experiment Ontology

★ The ontology comprehensively models the experimental details, such as processes, instruments, parameters, sample details, etc. along with Provenance (derivation history of results).

★ PEO has 118 classes and 27 properties with a logic expressivity of ALCHQ(D).



## References

<sup>1</sup>Trykikipedia homepage <http://wiki.knoesis.org/Trykikipedia>

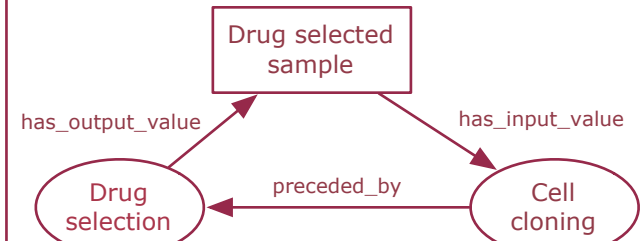
<sup>2</sup>National Center of Biological Ontology, <http://bioontology.org>

## Benefits of using PEO and PLO

**1. Integrating data sets:** PEO and PLO are used as schema to convert heterogeneous data that are in relational database, XML and CSV formats to RDF.

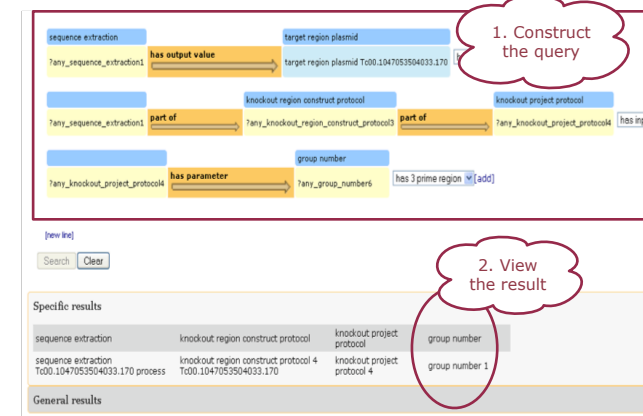
### 2. Reasoning over data set

Rule: If an entity is the output of process A and it is also the input of process B, then B is preceded by A.



### 3. Answering complex queries using Cuebee

Sample query : List all the groups that are using "Tc00.1047053504033.170" target region plasmid.



The screenshot shows the Cuebee query interface. The top part shows the query construction step: '1. Construct the query'. The bottom part shows the result view: '2. View the result'. The query is: 'List all the groups that are using "Tc00.1047053504033.170" target region plasmid.' The results table shows the following data:

sequence extraction	knockout region construct protocol	knockout project protocol	group number
sequence extraction Tc00.1047053504033.170 process	knockout region construct protocol 4 Tc00.1047053504033.170	knockout project protocol 4	group number 1